Re-ranking by Local Re-scoring for Video Indexing and Retrieval

Bahjat Safadi, Georges Quénot

Re-ranking by Local Re-Scoring for Video Indexing and Retrieval

Bahjat Safadi
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
Bahjat.Safadi@imag.fr

Georges Quénot
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
Georges.Quenot@imag.fr

ABSTRACT

Video retrieval can be done by ranking the samples according to their probability scores that were predicted by classifiers. It is often possible to improve the retrieval performance by re-ranking the samples. In this paper, we proposed a re-ranking method that improves the performance of semantic video indexing and retrieval, by re-evaluating the scores of the shots by the homogeneity and the nature of the video they belong to. Compared to previous works, the proposed method provides a framework for the re-ranking via the homogeneous distribution of video shots content in a temporal sequence. The experimental results showed that the proposed re-ranking method was able to improve the system performance by about 18% in average on the TRECVID 2010 semantic indexing task, videos collection with homogeneous contents. For TRECVID 2008, in the case of collections of videos with non-homogeneous contents, the system performance was improved by about 11-13%.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods, Indexing methods*; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

General Terms

Algorithms, Experimentation

Keywords

Video Indexing and Retrieval, Re-ranking

1. INTRODUCTION

Semantic indexing and retrieval for video databases has been a very active research field over the past few years. The global goal is to automatically describe the videos, then to index them by their contents. Nevertheless, retrieving relevant samples and easily navigating within large collection, are still very difficult tasks.

Generally, semantic indexing is achieved by supervised learning approaches, in which it is based on training classifier on positive and negative samples of a target concept (the development set). This classifier will generate a model, which will be used to predict the likeliness of new samples (the test set) to contain the target concept. The likeliness is often computed homogeneously to a probability for each data sample to contain the concept. Retrieval can then be done by ranking the samples according to their probability scores. Such ranking is initially done with a score independently for each sample using only information from the development set. It is often possible to improve the indexing or retrieval performance by re-ranking the samples, considering the results of the initial ranking on the whole test collection. Thus, re-ranking may lead to retrieve more relevant samples at the top of the ranked list. Recently, several methods have been proposed and developed for re-ranking. We review below some of these methods.

In context fusion [3, 5], the results of different searching models (concept-based search model, text-based search model and query by example) are used to re-rank the ranked lists. In fact, the focus here is on fusing output scores of different models. This method needs to train new classifiers on new descriptors. Since we also use, in our work, the fusion of output scores obtained by multiple models, we took this as a baseline approach.

Classification-based re-ranking [4], where the initial results of a baseline system are used to discover the co-occurrence patterns between the target semantics and extracted features. This is very similar to *"learning to rank"*[2], which is based on training a ranking model which can precisely predict the ranking lists in the dataset. In [4], the authors used the top-ranked and low-ranked samples respectively, as pseudo-positive and pseudo-negative examples to train a new classification model for ranking, and the classification margin for a target concept is regarded as its (new) re-ranked. The use of SVM as the classification model, leads to the method called *RankSVM*[2]. Ordinal re-ranking as it was proposed in [8], where the author re-ranks an initial results by using the co-occurrence patterns via the ranking functions. The final score is the weighting combination of the original score and the re-ranked scores. They adopted a training method to train the Re-ranking algorithm on some concepts, and the re-ranking algorithm was applied to re-rank the remaining concepts.

In the case of video collections: the retrieval units are often some video shots, rather than the whole videos themselves. Our contribution in this paper is to re-rank the video shots according to their initial scores, which were obtained from initial classifiers, and according to the video knowledge and nature. Compared with the work in [7], where the authors re-ranked the initial results of shots using the video knowledge score, which was estimated by calculating the arithmetic mean on the initial scores of all shots within the same video. This paper goes further: the generalized mean rule was adopted to calculate global score for each shot, depending on the knowledge obtained from the scores of its neighbors within the video, and it has been proved to be more efficient. Moreover, we studied the effectiveness of the re-ranking when applied on homogeneous and non homogeneous databases. Two windowing functions, the Rectangular and the Gaussian, were used on the neighbors of each shot to calculate its global score.

The paper is organized as follows: Our re-ranking method is presented in section 2, section 3 describes the experimental results and section 4 presents our concluding remarks.

2. RE-RANKING METHOD

In multimedia systems based video retrieval, we need to rank the video shots according to an estimation of their relevance to what the user is likely to want to see. This estimation can be the prediction score obtained by the trained model, as the likeliness of a shot to contain a target concept. Usually the order of the samples in the ranked lists contains some irrelevant samples, where in this case we can use a re-scoring method in order to minimize the error within these ranked lists.

The method we proposed here, considers the hypothesis that videos have rather homogeneous contents, and that the presence of a given concept in a video depends a lot on the nature of the video itself, and that the estimated scores are computed independently for all video shots in the corpus. The proposed re-ranking method is done by re-scoring the video shots, and this is done in two steps: First, for each shot, we compute global score z, this is calculated through the initial scores of its predefined neighbors within the same video. Then this global score will be used to re-evaluate the initial score of each shot. Let the test collection consists of a set of videos $V = (v_1, v_2, \ldots, v_m)$, m being the number of videos in the collection. Each video v_i composed of a sequence of shots $v_i = (s_{i1}, s_{i2}, \ldots, s_{in_i})$, n_i being the number of shots of v_i. For each shot s_{ij}, an initial classification score x_{ij} is computed from supervised learning on the development set.

Many options −including *(arithmetic mean, min, max, geometric mean, harmonic mean and root mean square)*− are possible for the computation of a global score z_{ij} for the shot x_{ik} in video v_i, from its neighbor shots. We considered the formula of a generalization mean rule, equation 1, to be the method to calculate the global scores of each shot in the video, since all the above methods can be inherited from this rule, by evaluating different parameters of α.

$$z_{ij} = \left(\frac{\sum_k f_\theta(j,k)(x_{ik})^\alpha}{\sum_k f_\theta(j,k)} \right)^{1/\alpha} \quad (1)$$

where x_{ik} indicates the score of shot k in video i, and α defines the used function, and it has to be tuned by cross-validation. Hence, different values of α leads to different functions, such as:($\alpha = -\infty$: Min; $\alpha = \infty$: Max; $\alpha = 0$: Geometric Mean; $\alpha = 1$: Arithmetic Mean; $\alpha = -1$: Harmonic Mean and $\alpha = 2$: Root Mean Square). $f_\theta(j,k)$ works as a window around the current shot j, to its neighbor shots in the $video_i$. In this paper, two kinds of windowing functions are considered: the rectangular ("hard") and the Gaussian ("soft"). In both cases, the size of the window is defined by a parameter θ. For the rectangular window, the number of neighbors of each shot in video i is given by $2\theta + 1$.For the Gaussian window, we have applied $\sigma = \sqrt{\theta(\theta+1)/3}$ so that both windowing functions have the same variance for the same value of θ. This θ parameters has also to be tuned within the training (or development set). $\theta = 0$ gives the baseline, it correspond to the initial ranking and $\theta = \infty$ uses a global score of the video itself which is calculated from all the shots belonging to it, in other words ($z_{ij} = z_i$).

After these global scores z_{ij} are calculated, the score of each shot is updated according to its previous score and its global score obtained from the video (its neighbors) knowledge. Again, many options were possible and we chose a weighted multiplicative fusion:

$$x'_{ij} = x_{ij}^{1-\gamma} \times z_{ij}^{\gamma}, \quad (2)$$

where γ is a parameter that controls the "strength" of the re-ranking method. This parameter also has to be tuned by cross-validation within the development collection.

3. EXPERIMENTS

In this section, we present our experiments in which we have evaluated the proposed re-ranking method on semantic indexing task. The experiments were conducted on TRECVID 2008 and 2010 databases. Each database consists of two large sets; the development and the test set. Table. 1 shows general information about these two databases. The TRECVID 2010 development set (2010d) consists of 119685 shots of 3173 videos with average of 37 shots per video, and the test set (2010t) consists of 146788 shots of 8467 videos with average of 17 shots per video, which might tells that videos in this database are homogeneous. The TRECVID 2008 development set (2008d) consists of 43616 shots of 219 videos with average of 199 shots per video, and the test set (2008t) consists of 42461 shots of 219 videos with average of 193 shots per video, videos are not homogeneous.

Table 1: databases' sizes.

Collection		Concepts	Shots/Videos	Min/Mean/Max
2010	dev	130	119685/3173	1/37/1381
	test	30	146788/8467	1/17/1423
2008	dev	20	43616/219	19/199/1003
	test	20	42461/219	14/193/1029

3.1 Re-ranking on semantic indexing task TRECVID 2010

This experiment was conducted on TRECVID 2010, which provided 130 concepts with ground truth labels in a training set. The evaluation was done by calculating the Mean Average Precision (MAP) on only 30 concepts that were chosen by NIST. We have evaluated the re-ranking method on four different initial classification results, which have been submitted to TRECVID 2010, including different fusion strategies such as weighted and direct optimized weighted fusion *(Fusion_MAP and Fusion_OPT)*, also the combination of

these two fusion types with the genetic fusion *(Fusion_GA_MAP and Fusion_GA_OPT)*. These fusion strategies were applied on score vectors obtained by training different systems on 45 different descriptors including audio and visual descriptors, which have been produced by the partners of the IRIM project of the GDR-ISIS [1]. Each of these fusion ways -of the classification results- can be considered as the context fusion method, which we took as the baseline method to our re-ranking algorithm.

3.1.1 Parameters' optimization

The tuning of α, θ and γ parameters (Eq. 1 and Eq. 2 in section 2), was conducted using the aforementioned initial classification results, which are calculated on the TRECVID 2010 development set. The aim of this tuning is to get the best values of α, θ and γ that gives the best performance of our system.

Figure 1 shows the results of tuning α and γ, in which we report the performance of the system in function with γ. We used the MAP on the 130 concepts as evaluation metric. Each plot, in the (top) sub-figure, is related to different value of α, and it shows the MAP with different values of γ (including $\gamma = 0$ and $\gamma = \text{inf}$). For each video, we have set θ to be the number of all shots related to it, and we have used the initial scores of *Fusion_MAP* for evaluation. From the plots, as we can see, $\alpha = 1$ and $\alpha = 2$ are performing better that the others, and the best result can be obtained with $\alpha = 2$ (Root Mean Square) and $\gamma = 0.4$. Moreover, in Figure 1 (bottom), we show the performance of the system, on the same collection, using the four used initial scores with $\alpha = 2$. As we can see, the highest performance, on each of the initial scores, was achieved when applying the re-ranking method with $\gamma = 0.4$.

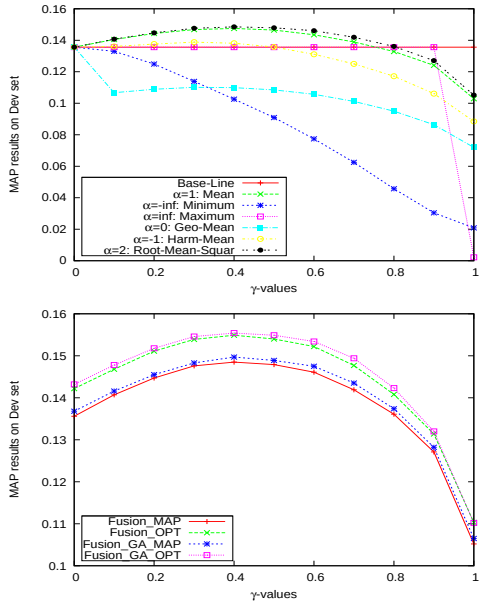

Figure 1: Tuning α (top) and γ (bottom) parameters on TRECVID 2010 development set.

Let's consider now the θ parameter in Equation 1. As mentioned before, this parameter controls the range on which we expect the video to have homogeneous content. The optimal value for this range is likely to depend upon the collection contents. We rerun the previous evaluations with different values of θ, including the baseline $\theta = 0$ and $\theta = \infty$ which means that the global score of each video is assigned to all the shots belonging to it ($z_{ij} = z_i$). Figure 2 shows the MAP calculated on the 130 concepts on the *Fusion_GA_OPT* run, which we consider as the best run as shown in figure 1 (bottom). The evaluations were done using the Rectangular and Gaussian windows with different θ-equivalent parameters for the re-ranking method. We have applied a sliding window of size $2\theta + 1$, as the neighbors of shot j using rectangular function, and $\sigma = \sqrt{\theta(\theta+1)/3}$ using Gaussian window. Thus, the two windowing functions have the same variance for the same value of θ. As we can see, the best result was obtained when $\theta = \infty$ for the two window functions. This is due to the fact that videos in the TRECVID 2010 collection are quite short (a few minutes in average), and they have homogeneous contents. Thus, local re-scoring does not perform better than global re-scoring.

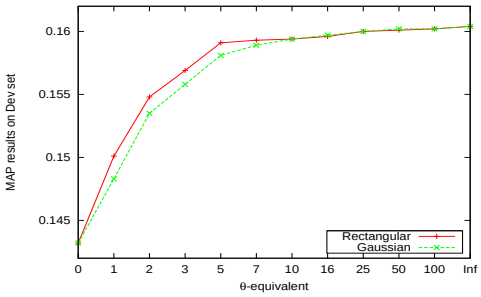

Figure 2: Tuning θ-equivalent parameters on TRECVID 2010 development set, using *Fusion_GA_OPT* run.

3.1.2 Evaluation on the test set

We have applied the proposed method on the TRECVID 2010 test set; with the best parameters obtained by the cross-validation, $\alpha = 2$, $\gamma = 0.4$ and $\theta = \infty$ with the two windowing functions (Rectangular and Gaussian. We have compared the new results −obtained after re-ranking− with the results of the initial scoring methods obtained using the best run; the *Fusion_GA_OPT* run.

We report, in table 2, the results (MAP on the 30 concepts) of the evaluation of the re-ranking method on TRECVID 2010 test set. As we can see, our proposed method has significantly improved the performance of the initial scoring methods; on this collection the proposed re-ranking method, with the fully homogeneity $\theta = \infty$, was able to improve the system performance with about 18% in average. The absolute MAP values are significantly different than in cross-validation (on the development set); this is mostly due to the fact that the set of concepts is different (30 only out of 130).

Table 2: Results of the re-ranking method on the test set of TRECVID 2010.

	θ/σ	MAP
Baseline	0	0.0480
ALL	∞	0.0568 (+18%)
Rectangular	$\theta = \infty$	0.0568 (+18%)
Gaussian	$\sigma = \infty$	0.0568 (+18%)

3.2 Re-ranking on HLF task TRECVID 2008

We have conducted the second experiment on TRECVID 2008 High-Level Feature Extraction task (HLF). Considering the Mean Average Precision (MAP) on these 20 con-

cepts to be the performance metric. The evaluation of the re-ranking method has been conducted using the simple late fusion of four types of image descriptors taken from IRIM GDR-ISIS partners [1], (including a combination of color histogram and Gabor transform, texture patterns, quaternionic wavelets and bag of SIFTs). The Multiple-SVM classifiers with RBF kernel was used as classification system, and it was implemented as in [6]. Since, TRECVID 2008 sets are not as homogeneous as TRECVID 2010 sets (see table. 1), we have fixed the optimal parameters $\alpha = 2$ and $\gamma = 0.4$, taken from section 3.1.1. The goal was to find the best value of θ for the re-ranking method, when dealing with non-homogeneous videos .

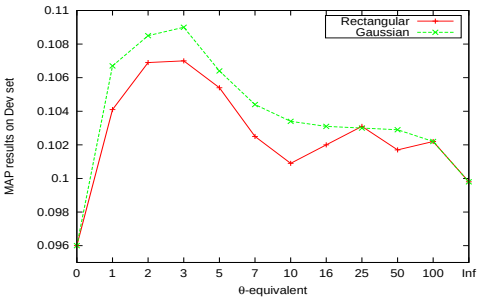

Figure 3: Tuning θ-equivalent parameter on TRECVID 2008, using the fusion of four descriptors.

We have evaluated our method on TRECVID 2008 development set using the late fusion of the four aforementioned descriptors, with different values of θ-equivalent parameter, within the same conditions as in section 3.1.1. We present the performance of the systems in Figure 3, which shows the MAP (calculated on the 20 concepts) with different values of θ-equivalent in both functions, the rectangular and Gaussian. As we can see, the Gaussian performance better than the rectangular function, and the performance using the two windowing functions was significantly enhanced when θ-equivalent is small and the best result was given when $\theta = 3$. In the Gaussian function when $\theta = 3 \rightarrow \sigma = \sqrt{3(3+1)/3} = 2$.

Furthermore, we have evaluated the re-ranking method with the optimal values $\alpha = 2, \gamma = 0.4$ and $\theta = 3$, on TRECVID 2008. It was evaluated using the two windowing functions. We report the final results in Table 3, in which we show the performance using deferent values of θ: $\theta = 0$ is the baseline, $\theta = \infty$ corresponds to applying the re-ranking on the whole videos, and the optimal θ-equivalent values(θ/σ) which defines respectively the rectangular and Gaussian functions. As we can see, the re-ranking with the optimal θ can significantly enhance the performance of the retrieval system. As expected, this collection is not homogeneous and there is not much enhancement when re-ranking by a global score on the whole video. When applying the re-ranking with $(\alpha = 0.4, \gamma = 0.4, \theta = 3)$, the performance of the system was enhanced in average by about 11-13% on the late fusion of the used descriptors with both the Gaussian and the rectangular windows.

Table 3: Results of the re-ranking method on the Test set of TRECVID 2008.

	θ/σ	MAP
Baseline	0	0.099
ALL	∞	0.101 (+2%)
Rectangular	$\theta = 3$	0.112 (+13%)
Gaussian	$\sigma = 2$	0.109 (+11%)

4. CONCLUSIONS

Video retrieval can be done by ranking the samples according to their probability scores that were predicted by classifiers. It is often possible to improve the retrieval performance by re-ranking the samples. In this paper, we proposed a re-ranking method that improves the performance of semantic video indexing and retrieval, by re-evaluating the scores of the shots using the homogeneity and the nature of the video they belong to.

The experimental results showed that the proposed re-ranking method was able to improve the system performance by about 18% in average on the TRECVID 2010 semantic indexing task, videos collection with homogeneous contents. For TRECVID 2008, in the case of collections of videos with non-homogeneous contents, the system performance was improved by about 11-13%.

Acknowledgements

This work was partly realized as part of the Quaero Program funded by OSEO, French State agency for innovation.

5. REFERENCES

[1] D. Gorisse, F. Precioso, P. Gosselin, L. Granjon, D. Pellerin, M. Rombaut, H. Bredin, L. Koenig, H. Lachambre, E. El Khoury, R. Vieux, B. Mansencal, Y. Zhou, J. Benois-Pineau, H. Jégou, S. Ayache, B. Safadi, Y. Tong, F. Thollard, G. Quénot, A. Benoit, and P. Lambert. IRIM at TRECVID 2010: High Level Feature Extraction and Instance Search. In *TREC Video Retrieval Evaluation workshop*, Gaithersburg, MD USA, nov 2010. NIST.

[2] R. Herbrich, T. Graepel, and K. Obermayer. Support vector learning for ordinal regression. In *Ninth Intl. Conf. on Artificial Neural Networks*, pages 97–102, 1999.

[3] W. Jiang, S.-F. Chang, and A. C. Loui. Context-based concept fusion with boosted conditional random fields. In *ICASSP (1)*, pages 949–952, 2007.

[4] L. S. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 333–340, New York, NY, USA, 2007. ACM.

[5] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li. Video search re-ranking via multi-graph propagation. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 208–217, New York, NY, USA, 2007. ACM.

[6] B. Safadi and G. Quénot. Active learning with multiple classifiers for multimedia indexing. Grenoble, France, June 2010. CBMI.

[7] F. Wang and B. Merialdo. Eurecom at trecvid 2009 high-level feature extraction. In *TREC2009 notebook*, 16-17 Nov 2009.

[8] Y.-H. Yang and W. H. Hsu. Video search reranking via online ordinal reranking. In *ICME*, pages 285–288, 2008.